iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59156

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Air Quality Index Prediction Using DL

Srinivas Rao Pendela[1], Vakula Sai Bhavani Myla[2], Bhanu Prakash Meduri[3], Pavan Kumar Palem[4], Harshavardhan Reddy Mutyam[5]

[1]*Associate Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology (Autonomous), Guntur, AP*
[2,3,4,5] *UG Students, Department of CSE, Vasireddy Venkatadri Institute of Technology (Autonomous), Guntur, AP*

Abstract: ***Air pollution emerged as greatest environmental threat to human health and the planet. Occurred due to release of harmful substances like $PM_{2.5}$, PM_{10}, oxides of nitrogen, carbon and sulphur, ozone, volatile organic compounds etc.. into air by various human activities like vehicle exhaust, agricultural practices, cooking, combustion, burning of fossil fuels, use of air conditioners, refrigerators and other sources adversely effecting climate, ecosystems, health and biodiversity. Assessing and monitoring air pollution levels became obligatory. The current study deals with air pollution data collected from various cities in India over a period of six years (2015-2020). Deep learning models LSTM and GRU are employed to predict the air quality index. The dataset underwent meticulous preprocessing, followed by a sophisticated feature selection process utilizing Principal Component Analysis (PCA). This strategic approach allowed us to discern and isolate the principal pollutants exerting a substantial influence on air quality. The application of LSTM and GRU models for AQI prediction holds great potential in improving our understanding of air pollution dynamics.***

Keywords: Air pollution, Deep Learning, LSTM, GRU, Air Quality Index, Feature Selection, Principal Component Analysis.

I. INTRODUCTION

In the contemporary era, the escalating concerns about air quality have reached a critical juncture, reflecting a pressing global challenge. The ubiquitous presence of pollutants in the atmosphere, stemming from a myriad of sources such as industrial emissions, vehicular activities, and urbanization, has unleashed a cascade of adverse effects on both human health and the environment. Particulate matter, nitrogen dioxide, sulphur dioxide, and an array of pollutants collectively contribute to a spectrum of health issues, ranging from respiratory ailments to cardiovascular complications. As urban landscapes burgeon and industrial activities intensify, the rapid surge in air pollution has become an alarming phenomenon. This surge has become a significant catalyst for various health risks, necessitating an in-depth understanding and predictive measures to mitigate its impact.

In recent years, the consequences of deteriorating air quality have manifested in headline-grabbing events across the globe, underlining the urgency of addressing this environmental crisis. From unprecedented smog episodes engulfing major cities to severe health repercussions affecting vulnerable populations, the news has been rife with alarming incidents.

For instance, cities like Beijing and New Delhi have made international headlines due to notorious "airpocalypses," where dense smog shrouded the urban landscapes, leading to public health emergencies. These events have not only resulted in a surge of respiratory illnesses but have also prompted governments to implement stringent measures to curb pollution. Moreover, incidents such as the Australian bushfires have underscored the interconnectedness of air quality and natural disasters. The massive scale of these fires not only emitted colossal amounts of pollutants into the atmosphere but also had far-reaching implications for air quality, posing serious health risks for communities in the vicinity and beyond. These noteworthy occurrences serve as poignant reminders of the far-reaching implications of compromised air quality, motivating concerted efforts to leverage advanced technologies and methodologies for accurate prediction and mitigation. Air quality index can be categorized into six types as shown in the figure below.

AQI	Remark	Color Code	Possible Health Impacts
0-50	Good		Minimal impact
51-100	Satisfactory		Minor breathing discomfort to sensitive people
101-200	Moderate		Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor		Breathing discomfort to most people on prolonged exposure
301-400	Very Poor		Respiratory illness on prolonged exposure
401-500	Severe		Affects healthy people and seriously impacts those with existing diseases

Fig.1 Classification of AQI into 6 categories

In this context, the application of advanced technologies like deep learning has gained prominence for predicting the Air Quality Index (AQI). Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), as formidable components of deep learning, offer a potent solution for deciphering the intricate patterns inherent in time-series air quality data. In the realm of air quality prediction, addressing missing values is a crucial aspect to ensure the reliability of the models and insights derived from the data. One effective technique employed for filling these gaps is interpolation. Interpolation serves as a valuable method for estimating the values of missing data points based on the information available from neighbouring observations. Moreover, navigating the complex landscape of air quality prediction requires judicious selection of influential pollutants. Principal Component Analysis (PCA) emerges as a valuable ally, aiding in the identification of major contributors to air quality fluctuations.

This research delves into the realm of air quality prediction, harnessing the capabilities of LSTM and GRU while utilizing PCA for feature selection. By addressing challenges related to data imbalance and missing values, this study aims to unravel nuanced insights that can significantly contribute to the comprehension and management of the escalating air pollution crisis.

II. LITERATURE REVIEW

The comprehensive study conducted by D. Kothandaraman, N. Praveena et al. (2022)[1] aimed at predicting PM2.5 pollutant levels and assessing air quality using datasets acquired from Anand Vihar, Delhi, courtesy of the Delhi Pollution Control Committee. This dataset spans from January 1, 2014, to December 1, 2019, capturing continuous 24-hour monitoring of PM2.5 levels, alongside meteorological features. The proposed models encompass linear regression, random forest, KNN, ridge and lasso regression, XGBoost, and AdaBoost. The evaluation of these models involved statistical metrics, including MAE, MAPE, MSE, RMSE, and R^2. Notably, the XGBoost, AdaBoost, and random forest models emerged as reliable choices, exhibiting low MAE, MAPE and RMSE values, respectively.

Madhuri VM et al. (2020)[2], the research considered a variety of parameters, including CO, Tin oxide, non-metallic hydrocarbons, Benzene, Titanium, NO, Tungsten, Indium oxide, and Temperature. Linear Regression, SVM, DT, and RF were employed to predict the Relative Humidity of air. The evaluation metric RMSE, was utilized to gauge accuracy. The findings revealed that the AQI predictions obtained through Random Forest method were promising, showcasing its efficacy in predicting air quality with notable accuracy, as discussed in the analysed results.

In a study conducted by Yves Rybarczyk and Rasa Zalakeviciute (2021)[3], the impact of the COVID-19 outbreak on air quality in Quito, Ecuador, was assessed using ML models based on a Gradient Boosting Machine algorithm. The precision of the predictions was initially evaluated through cross-validation over four years prelockdown, demonstrating high accuracy in estimating actual pollution levels. Subsequently, the study quantified changes in pollution during the full lockdown, revealing significant decreases of approximately $-53 \pm 2\%$, $-45 \pm 11\%$, $-30 \pm 13\%$, and $-15 \pm 9\%$ for NO2, SO_2, CO, and $PM_{2.5}$, respectively. The most impacted areas were identified as traffic-busy districts within the city. Following the transition to partial relaxation, pollution concentrations almost returned to pre-pandemic levels. The quantification of the pollution drop was supported by an assessment of the prediction's mean Root RMSE and mean Pearson Correlation Coefficient (PCC), further validating the study's findings.

Doreswamy et al. (2020)[4], ML predictive models were explored for forecasting concentrations of particulate matter in atmospheric air using Taiwan Air Quality Monitoring data sets spanning from 2012 to 2017. Performance metrics include RMSE (Root Mean Square Error), MAE (Mean Absolute Error), MSE (Mean Square Error), and R^2 were employed. ML algorithms such as linear regression, random forest regressor, gradient boosting regressor, k neighbours regressor, MLP regressor, and Decision Tree regressor CART were utilized for forecasting. The study concluded, gradient boosting regressor model outperformed others in predicting air pollution on the TAQMN data, highlighting its effectiveness in this context.

Ling Qing's (2023)[5], introduces a deep learning method using the GRA-GRU model for regional $PM_{2.5}$ concentration prediction in Changchun City. The approach incorporates various meteorological elements and employs grey relational analysis, constructing a spatial weight matrix to capture spatial relationships. The GRU model, chosen for its simplicity and efficiency over traditional LSTM models, contributes to improved accuracy in predicting $_{PM2.5}$ concentrations and simulating temporal changes in air pollutant levels for the specified region.

Nilesh N. Maltare et al. (2023)[6], the authors investigated the effectiveness of SARIMA, SVM with various kernel functions, and LSTM models for predicting AQI. Evaluation metrics such as the Coefficient of Determination (R^2), MSE, and RMSE are employed to assess model performance. SVM with the RBF kernel outperformed the other models, suggesting its potential as a reliable method for AQI prediction in Ahmedabad city.

Ruiyun Yu et a. (2016)[7], the authors compared the performance of several machine learning methods, including Naïve Bayes, Logistic Regression, Single Decision Tree, ANN, and their proposed Random Forest approach. The evaluation was based on various measurements such as precision, recall, F-score, Relative Absolute Error (RAE), and Receiver Operating Characteristic (ROC). The results demonstrated that RAQ outperformed the other methods, indicating its efficacy in predicting air quality in urban sensing systems.

Nahar K et al. (2020)[8], various ML algorithms were employed to predict air quality in Jordan. The models tested included DT, SVM, k-Nearest Neighbor (k-NN), Random Forest, and Logistic Regression. The dataset used in the study covered the period from January 1, 2017, to April 30, 2019, and included 9777 records from 13 different locations in Jordan. All algorithms demonstrated good accuracy in predicting air quality, with Decision Tree showing the highest performance among them. These results suggest that Decision Tree may be a suitable choice for air quality prediction in Jordan based on the dataset and timeframe analysed.

Varsha Gopalakrishnan (2021)[9], presents a novel approach to predict concentration levels of nitrogen oxides and black carbon in hyperlocal environments. The study leverages Google Street View data as input to machine learning models, complemented by additional features such as meteorological data and traffic data. The author seeks to enhance the precision of air quality forecasts at a detailed, hyperlocal level by combining various datasets. From the five models tested, such as XGBoost and Random Forest, these two stand out as the most proficient in anticipating air pollutant levels. The study provides valuable insights into the promise of machine learning methods for hyperlocal air quality forecasting, underscoring the significance of integrating diverse data sets for improved predictive accuracy. This work contributes to the advancement of environmental monitoring and management strategies, with implications for public health and urban planning.

Mauro Castelli et al. (2020)[10], the author's utilized the Environmental Protection Agency's (EPA) dataset from California to develop forecasting models for five different pollutants: CO, NO2, SO2, ozone, and PM2.5. They employed SVR models, with one variant incorporating PCA for feature reduction, termed PCA SVR-RBF. The objective of their research was to predict air quality parameters by taking pollutant concentrations as input and outputting AQI. The study concluded that both models demonstrated good performance in predicting air quality parameters. However, SVR-RBF slightly outperformed PCA SVR-RBF, indicating the efficacy of the SVR model, particularly when combined with a radial basis function kernel. This research contributes valuable insights into the application of machine learning techniques for air quality prediction, providing a basis for enhancing environmental monitoring and management efforts.

Dyuthi Sanjeev (2021)[11], explored the effectiveness of RF, SVM, and ANN for predicting air quality. Using a dataset containing concentrations of pollutants and meteorological factors, the models were trained and evaluated. Random Forest-based model achieved the highest accuracy among the three algorithms, followed by the SVM and the Artificial Neural Network. These findings suggest that the Random Forest Algorithm is the most efficient in this context.

Ghufran Isam Drewil et al. (2022)[12], the author's focused on forecasting the individual values of PM2.5, PM10, CO, and NOx pollutants for the following day. Utilizing a dataset similar to the one being analysed, the authors employed a LSTM model integrated with Genetic Algorithm (GA) for optimization. The primary objective was to predict the levels of air pollution for a group of pollutants, namely PM2.5, PM10, CO, and NOx, with the LSTM model enhanced by GA exhibiting superior performance compared to other variants, including bi-LSTM and c-LSTM. This finding underscores the effectiveness of leveraging metaheuristic algorithms like GA to optimize deep learning models for air quality prediction. The study provides valuable insights for advancing precise forecasting models essential for environmental monitoring and public health management.

Jingyang Wang et al. (2022)[13], AQI is predicted using a combination of CNN for feature selection and LSTM for AQI prediction. Their dataset comprised six pollutants: PM2.5, PM10, CO, O3, NO2, and SO2, with AQI as the target label, focusing specifically on forecasting AQI levels for Shijiazhuang City in China. The authors compared their CNN-ILSTM model with traditional regression models such as SVR, Random Forest Regression (RFR), and Multilayer Perceptron (MLP), as well as various deep learning models including LSTM, GRU, Integrated LSTM (ILSTM), CNN-LSTM, and CNN-GRU. The overall evaluation of prediction results indicated that CNN-ILSTM outperformed all other models considered. This finding underscores the effectiveness of integrating CNN for feature selection with LSTM for AQI prediction, providing valuable insights into the advancement of air quality prediction methodologies.

Tanisha Madan et al. (2020)[14], author's utilized a Kaggle dataset to predict AQI. The study employs various ML algorithms, including Linear Regression, DT, Random Forest, ANN, Support Vector Machine (SVM), boosting, and neural networks, to forecast AQI levels. However, the authors do not provide specific implementation details in their review. Despite this, their research concludes that among the evaluated algorithms, Linear Regression, Decision Tree, Random Forest, ANN, SVM, boosting, and neural networks demonstrate promising results for air quality prediction.

This study contributes to the understanding of machine learning techniques for AQI forecasting, offering insights into the potential effectiveness of different algorithms in addressing environmental challenges.

Abdellatif Bekkar et al. (2021)[15], author's employed advanced deep learning models, including LSTM, GRU, and CNN, along with hybrid architectures such as CNN-LSTM, to predict the hourly forecast of PM2.5 concentration in Beijing, China.

Utilizing a dataset sourced from the UCI Machine Learning Repository, the study integrated pollutant components and meteorological data to explore the correlation between features and PM2.5 levels. Their experimental analysis compared the performance of various deep learning algorithms, including LSTM, Bi-LSTM, GRU, Bi-GRU, CNN, and the hybrid CNN-LSTM model. The results indicated that the hybrid CNN-LSTM multivariate approach outperformed all other traditional models, demonstrating superior predictive performance. This research contributes valuable insights into the effectiveness of deep learning techniques for air pollution prediction, particularly in urban environments, paving the way for improved environmental monitoring and management strategies.

III.METHODOLOGIES

This paper proposes the use of LSTM and GRU methods for predicting air quality index. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are popular deep learning architectures used for time series forecasting. To employ LSTM and GRU for AQI prediction, historical data such as pollutant concentrations are used as input sequences. These models learn the patterns and trends from the historical data to predict future AQI values. LSTM and GRU are particularly beneficial for this task because they can handle long sequences of data, retain information over long periods. By training LSTM and GRU models on historical AQI and related data, they can learn the complex relationships between various factors and AQI levels. This enables them to make accurate predictions, which can be valuable for air quality monitoring and management.

A. Model Architecture

Our proposed model architecture for time series forecasting tasks is designed to effectively capture complex temporal dependencies inherent in sequential data. The architecture comprises two layers of LSTM units for LSTM model and two layers of GRU units for GRU model and one layer of dense, strategically arranged to exploit the hierarchical structure of time series data. This architecture is common for both LSTM and GRU models.

1) First Layer (64 Units with return_sequences=True)

The first layer serves as the foundation of our hierarchical architecture, consisting of 64 units.

Setting the parameter return_sequences to true allows the layer to retain and propagate sequential information to subsequent layers.

This configuration enables the model to capture fine-grained temporal patterns and dependencies within the input sequence.

By preserving the sequential nature of the data, the first layer facilitates the extraction of meaningful features that are essential for accurate forecasting.

2) Second Layer (32 Units)

The second layer builds upon the representations learned by the first layer, further refining the temporal features extracted from the input sequence.

With 32 units, this layer focuses on capturing higher-level temporal dependencies and patterns.

The hierarchical structure encourages the model to learn increasingly abstract representations of the input data, enhancing its ability to discern relevant information for forecasting purposes.

3) Dense layer (Output Layer consists 8 Outputs)

The final layer of the architecture produces 8 output predictions corresponding to the future time steps of the input sequence.

These predictions encapsulate the model's forecasted values for the time series at specific points in the future.

By aggregating information learned from the preceding layers, the output layer generates accurate and informative forecasts, aiding decision-making processes in diverse domains.

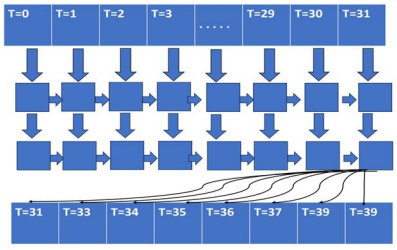


Fig. 2 Architecture

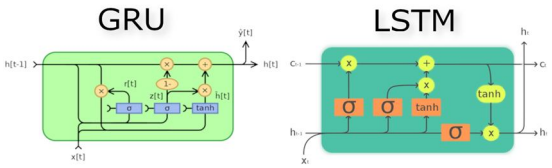


Fig. 3 GRU and LSTM Cell

B. LSTM Cell

1) Inputs

Input Sequence: The LSTM cell receives an input sequence, typically represented as a vector or a sequence of vectors, containing temporal information.

Previous Hidden State (ht-1): The hidden state from the previous time step serves as contextual information for the current time step.

Previous Cell State (ct-1): The cell state from the previous time step carries memory information from earlier time steps.

2) Outputs

New Hidden State (ht): The LSTM cell produces a new hidden state, which encapsulates the updated contextual information based on the input sequence and the previous hidden state.

New Cell State (ct): The LSTM cell generates a new cell state, which integrates both new information from the input sequence and relevant information retained from the previous cell state.

3) Gates

Forget Gate: Determines which information from the previous cell state to discard, based on the input sequence and the previous hidden state.

Input Gate: Determines which new information from the input sequence to incorporate into the cell state.

Output Gate: Controls which information from the current cell state to output as the hidden state for the current time step.

C. GRU Cell

1) Inputs

Input Sequence: Similar to the LSTM cell, the GRU cell accepts an input sequence containing temporal information.

Previous Hidden State (ht-1): The hidden state from the previous time step provides context for the current time step.

2) Outputs

New Hidden State (ht): The GRU cell produces a new hidden state, which represents updated contextual information based on the input sequence and the previous hidden state.

3) Gates

Reset Gate: Determines how much of the previous hidden state to forget, allowing the model to adaptively reset its internal state based on the input sequence.

Update Gate: Controls how much of the new information from the input sequence to incorporate into the hidden state, as well as how much of the previous hidden state to retain.

IV. IMPLEMENTATION

A. Data Cleaning

In the context of air quality index (AQI) prediction, the following data cleaning steps can be employed:

1) *Interpolation using Cubic Spline:* This technique involves hourly filling of missing values in the time series data by employing cubic spline interpolation. Cubic spline interpolation is a mathematical method for constructing a smooth curve that passes through the given data points. Specifically, missing values in the time series are replaced by estimated values obtained through cubic spline interpolation on an hourly basis. This process ensures that the resulting filled values maintain the continuity of the time series and preserve its overall shape and trends, effectively reconstructing the missing data points while accounting for neighboring observations within each hour of the day.
2) *Mean Imputation by Month:* In this method, missing values are filled by taking the mean of the available data for the corresponding month. This approach leverages the temporal structure of the data by using the average value for the specific month across multiple years. By replacing missing values with the mean of the month, it maintains the seasonality and monthly patterns present in the time series.
3) *Mean Imputation by Month-Day-Hour:* This technique involves filling missing values by computing the mean of the available data based on the combination of year, month, and hour. By grouping the data according to these three dimensions, it captures both the seasonal patterns and diurnal variations in the time series. Filling missing values with the mean of the corresponding month-day-hour group ensures that the imputed values are more localized and representative of the specific time within the temporal hierarchy.

These data cleaning steps help ensure that the time series data used for AQI prediction is complete, consistent, and representative, which can lead to more accurate and reliable predictions.

B. Feature Selection

Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction. It works by transforming the original features of a dataset into a new set of linearly uncorrelated variables called principal components. These components are ordered in such a way that the first component explains the maximum variance in the data, the second component explains the maximum remaining variance, and so on.

PCA is particularly useful when dealing with high-dimensional datasets, as it can help reduce the number of features while retaining most of the important information.

PCA is applied to a dataset containing air quality data. The dataset is first preprocessed to handle missing values and standardize the numeric features using the SimpleImputer and StandardScaler classes from the scikit-learn library, respectively. The standardized numeric features are then combined with non-numeric features (if needed) and used as input for PCA. The number of components to retain in PCA is specified as 12, which is the minimum of the number of samples in the dataset and the number of features. This ensures that PCA does not attempt to extract more components than there are observations or features, which could lead to overfitting.

After fitting the PCA model to the standardized data, the explained variance ratio for each principal component is calculated. The explained variance ratio indicates the proportion of the total variance in the dataset that is explained by each principal component. This information is useful for understanding how much information is retained by each component and can help in deciding how many components to retain for dimensionality reduction.

Finally, the explained variance ratio is visualized using a bar plot, with the pollutants or features on the x-axis and the explained variance ratio on the y-axis.

This plot provides a visual representation of how much each pollutant contributes to the overall variance in the dataset, helping to identify which pollutants are most important in explaining the variability in air quality data.

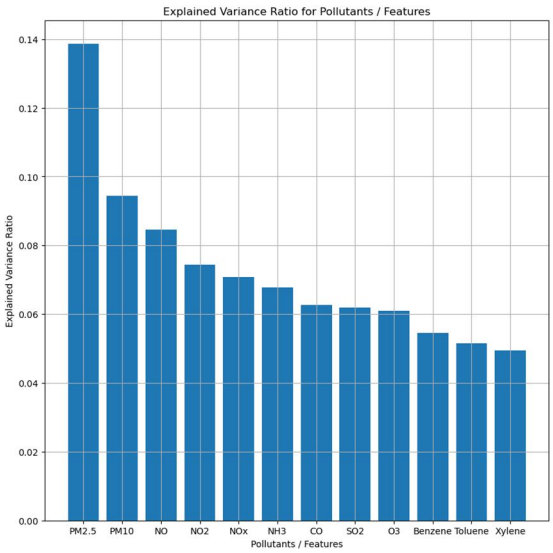


Fig. 4 Graphical representation - Results of PCA

Principal Component	Explained Variance Ratio
PM2.5	0.13855
PM10	0.0943418
NO	0.0846454
NO2	0.0743365
NOx	0.0708236
NH3	0.0678602
CO	0.0625961
SO2	0.061979
O3	0.0609872
Benzene	0.0546375
Toluene	0.0515654
Xylene	0.0493878

Fig. 5 Results of PCA

Both the LSTM and GRU models are constructed using the Keras library with TensorFlow backend. They consist of two recurrent layers followed by a dropout layer for regularization and a dense layer for output. Both models start with a recurrent layer with 64 units, followed by another recurrent layer with 32 units. The first layer in each model is set to return sequences to feed into the subsequent layer, while the second layer returns only the final output. A dropout layer with a dropout rate of 0.2 is added to prevent overfitting. Finally, a dense layer with 8 units is used for the output prediction. The models are compiled using the Adam optimizer and Mean Squared Error (MSE) loss function, suitable for regression tasks. For the LSTM model, training is carried out for 10 epochs with a batch size of 30, while for the GRU model, training is carried out with a batch size of 100.

V. RESULTS

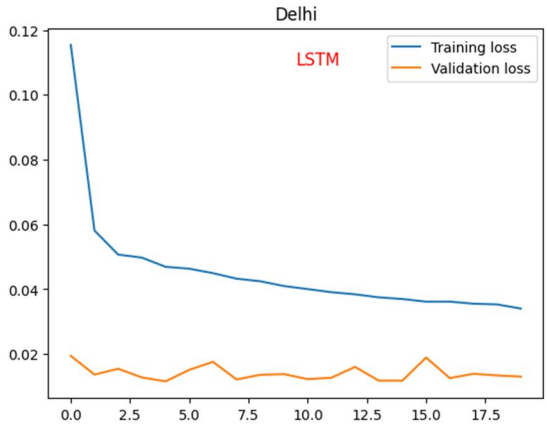


Fig.6 Training and Validation loss Using LSTM (Delhi)

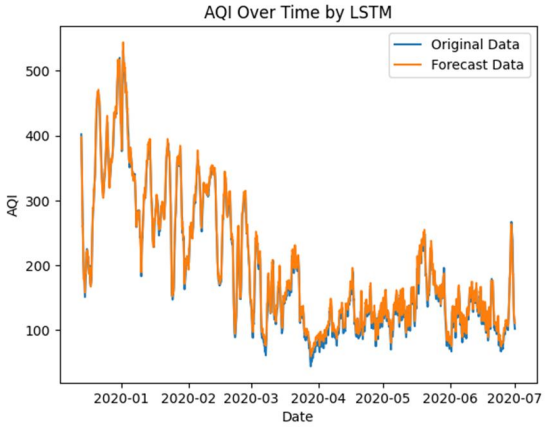


Fig.7 AQI Using LSTM (Delhi)

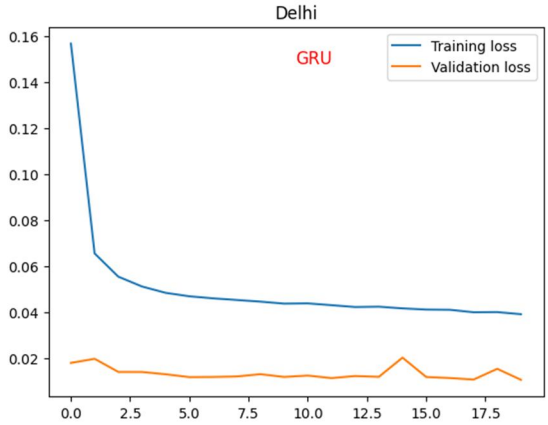


Fig.8 Training and Validation loss Using GRU (Delhi)

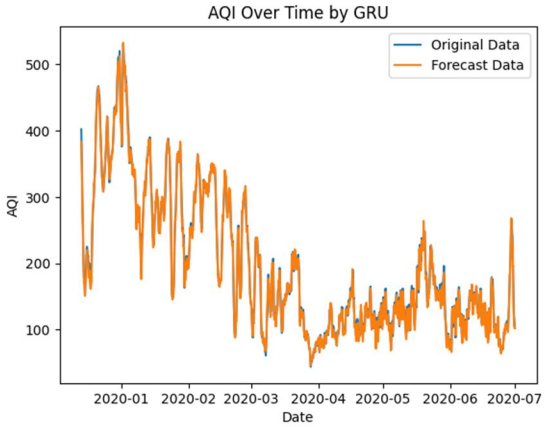


Fig.9 AQI Using GRU (Delhi)

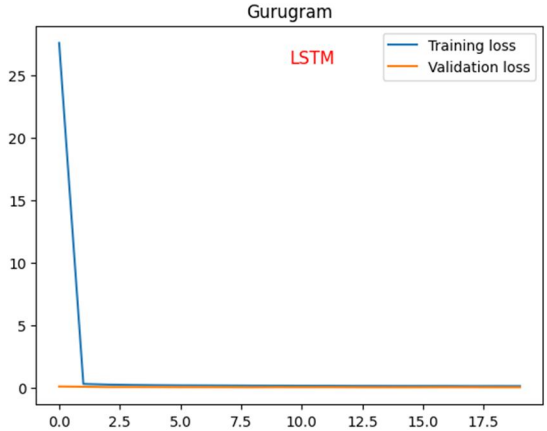


Fig.10 Training and Validation loss Using LSTM (Gurugram)

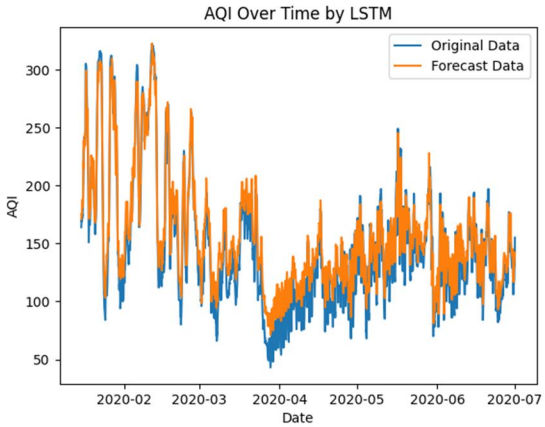


Fig.11 AQI Using LSTM (Gurugram)

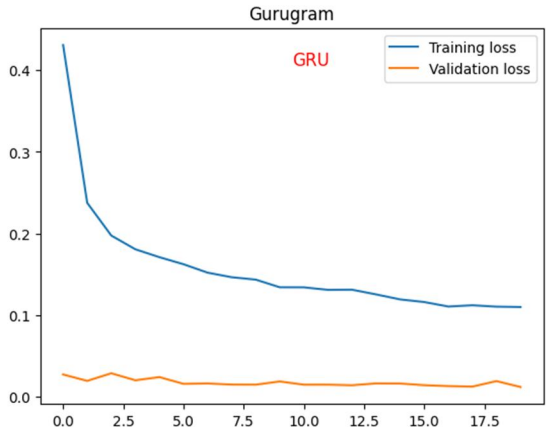


Fig.12 Training and Validation loss Using GRU (Gurugram)

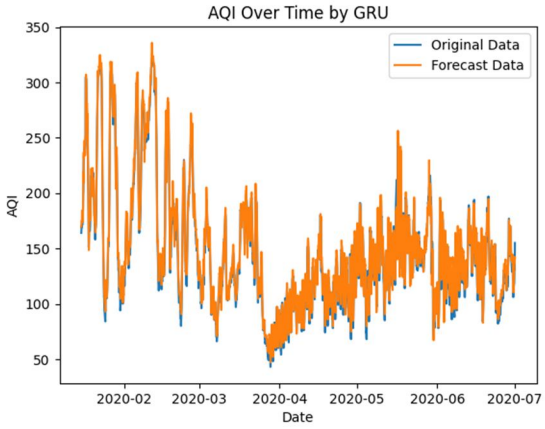


Fig.13 AQI Using GRU (Gurugram)

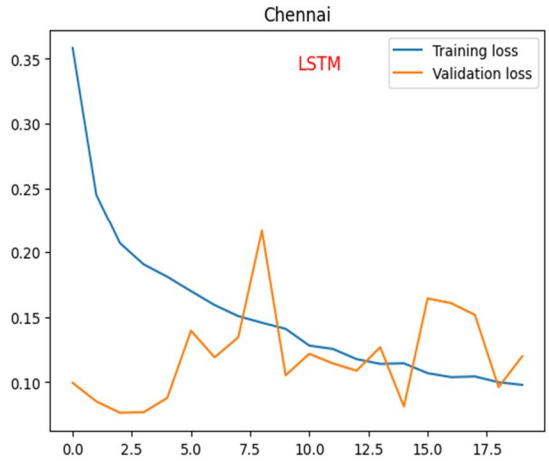


Fig.14 Training and Validation loss Using LSTM (Chennai)

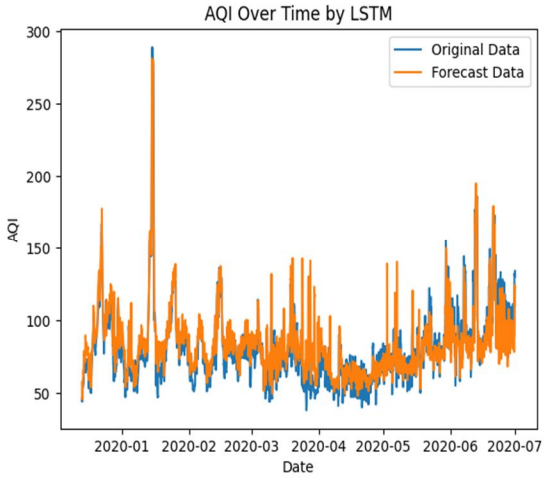


Fig.15 AQI Using LSTM (Chennai)

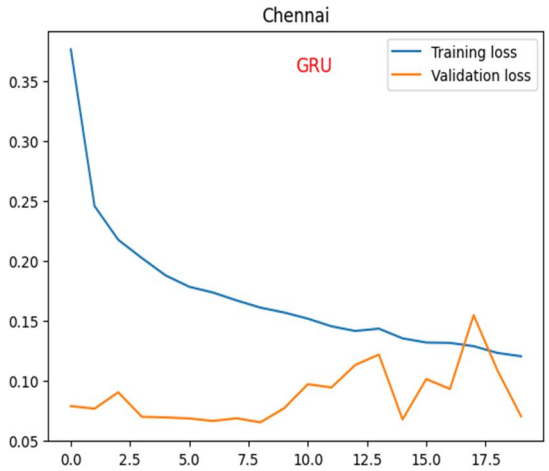


Fig.16 Training and Validation loss Using GRU (Chennai)

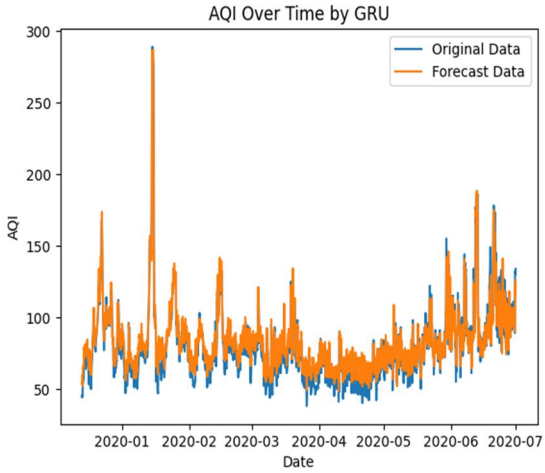


Fig.17 AQI Using GRU (Chennai)

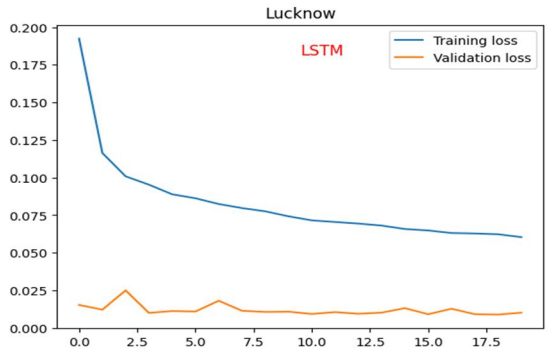


Fig.18 Training and Validation loss Using LSTM (Lucknow)

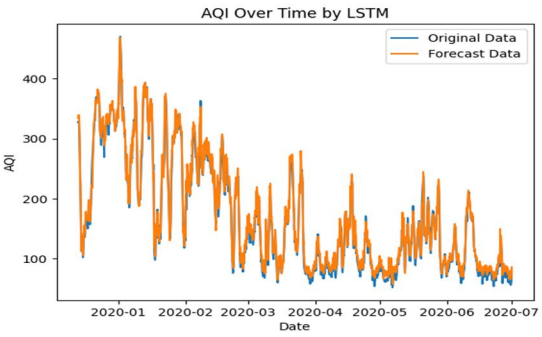


Fig.19 AQI Using LSTM (Lucknow)

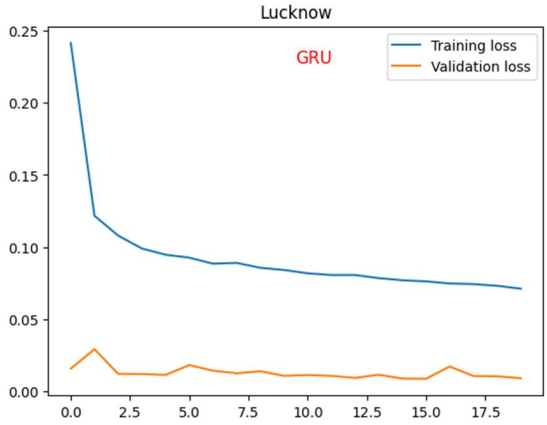


Fig.20 Training and Validation loss Using GRU (Lucknow)

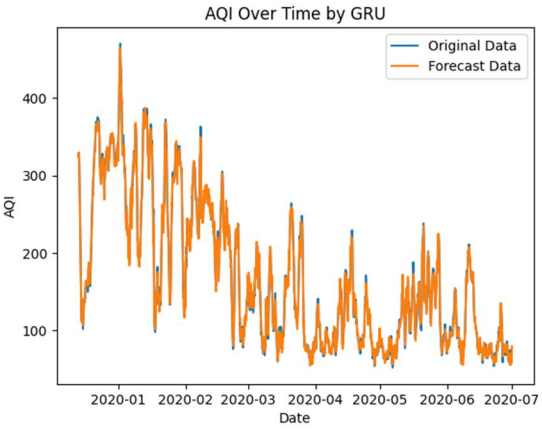


Fig.21 AQI Using GRU (Lucknow)

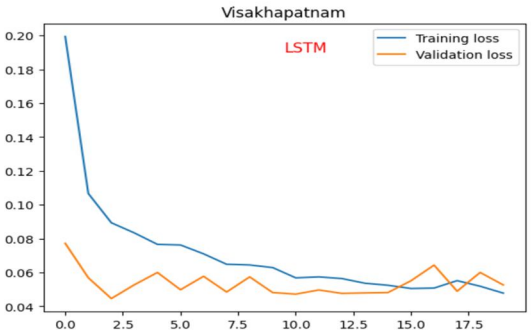


Fig.22 Training and Validation loss Using LSTM (Visakhapatnam)

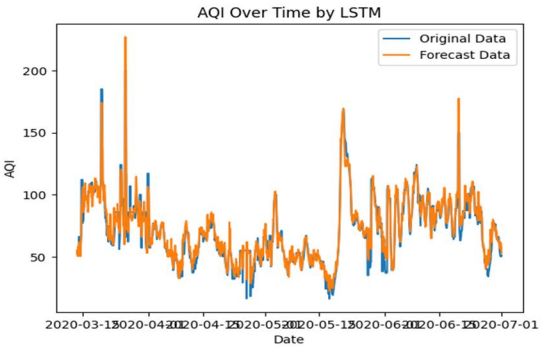


Fig.23 AQI Using LSTM (Visakhapatnam)

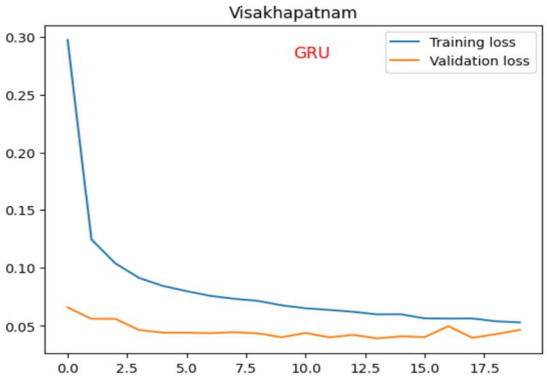


Fig.24 Training and Validation loss Using GRU (Visakhapatnam)

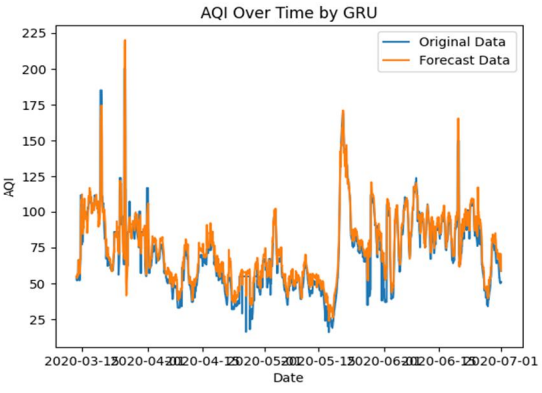


Fig.25 AQI Using GRU (Visakhapatnam)

S.No	City	LSTM (MAE)	LSTM (MSE)	LSTM (RMSE)	GRU (MAE)	GRU (MSE)	GRU (RMSE)
1	Aizwal	17.51	80.12	8.95	25.25	120.00	10.95
2	Amaravati	130.69	6436.05	80.22	73.83	2068.63	45.48
3	Amritsar	79.56	3297.35	57.42	80.90	3297.48	57.42
4	Bengaluru	153.46	4260.33	65.27	62.51	1155.20	33.99
5	Bhopal	104.77	2396.41	48.95	83.57	1546.67	39.33
6	Chandigarh	106.20	4267.68	65.33	99.98	2220.30	47.12
7	Chennai	97.81	2997.43	54.75	72.71	1273.77	35.69
8	Delhi	68.89	1056.51	32.50	63.52	973.83	31.21
9	Ernakulam	104.28	2936.66	54.19	68.92	2557.18	50.57
10	Gurugram	109.51	2521.83	50.22	93.93	2180.81	46.70
11	Guwahati	55.08	613.86	24.78	75.40	1015.48	31.87
12	Hyderabad	54.45	628.04	25.06	57.18	680.91	26.09
13	Jaipur	72.99	1376.15	37.10	60.34	968.34	31.12
14	Kolkata	68.45	886.47	29.77	58.58	796.27	28.22
15	Lucknow	103.86	2386.57	48.85	99.88	2867.39	53.55
16	Mumbai	29.92	203.99	14.28	23.30	141.30	11.89
17	Shillong	100.05	2857.20	53.45	90.98	3403.38	58.34
18	Talcher	352.43	29977.19	173.14	351.46	28707.64	169.43
19	Thiruvananthapuram	91.97	1705.52	41.30	57.27	780.66	27.94
20	Visakhapatnam	64.89	1393.20	37.33	66.76	1271.19	35.65

Fig.26 MAE, MSE, RMSE values for both models

The results of the project showcasing GRU's superior performance over LSTM in predicting AQI are not only promising but also offer valuable insights into the application of deep learning for time series analysis. The evaluation metrics, including MSE and RMSE, MAE consistently favoured the GRU model, indicating its ability to make more accurate predictions. This outcome underscores the significance of selecting the right model architecture and hyperparameters for time series forecasting tasks. The hyperparameters and the specific configuration of the models could have influenced their performance. Parameters like the number of units in each layer, the dropout rate, and the batch size can significantly impact the model's ability to learn and generalize from the data. Furthermore, the validation of GRU's effectiveness in handling complex temporal dependencies and shorter sequences suggests its potential for real-world applications where computational efficiency and accurate predictions are paramount. GRU is generally considered computationally more efficient than LSTM due to its simpler structure. GRU has two gates (update and reset gates) compared to LSTM's three gates (input, output, and forget gates), which could lead to faster convergence and better generalization, especially in cases where the data does not require complex memory management.

VI. CONCLUSIONS

In conclusion, this project aimed to predict the Air Quality Index (AQI), a crucial metric for assessing air quality and its impact on health and the environment. The need for such a project is evident due to the increasing concern over air pollution and its adverse effects on human health and the environment. By employing LSTM and GRU models, both known for their effectiveness in handling sequential data, along with PCA analysis for feature selection, the project sought to enhance the accuracy of AQI predictions. The results revealed that the GRU model outperformed the LSTM model, achieving lower MSE and RMSE values. This outcome underscores the importance of choosing the right model architecture and techniques like PCA for feature selection in improving the accuracy of AQI predictions. The success of the GRU model highlights its potential for real-world applications, offering a promising approach for monitoring and managing air quality to mitigate its detrimental effects.

VII. LIMITATIONS

One of the primary limitations of this project is the presence of inadequate data for some cities, leading to too many missing values. This limitation can significantly impact the accuracy and reliability of the predictions, as the models may not have enough data to learn and generalize effectively. Additionally, the quality of the data, including issues such as inaccuracies or inconsistencies, can also pose challenges to the modelling process and result in less reliable predictions.

Other limitations of the project include the assumption of linear relationships between features and the target variable, which may not always hold true in real-world scenarios.

Moreover, the use of historical data for predictions may not account for sudden changes or anomalies in air quality, limiting the models' ability to provide timely and accurate forecasts. Furthermore, the project's reliance on a specific set of features for prediction may overlook other potential factors influencing air quality, such as weather patterns, industrial activities, and traffic conditions. Incorporating these additional factors could enhance the models' predictive capabilities but would require more comprehensive and diverse datasets.

VIII. FUTURE WORK

In future work, integrating meteorological data into the existing models could significantly enhance the accuracy and robustness of air quality predictions. Meteorological factors such as temperature, humidity, wind speed, and precipitation can have a profound impact on air quality, influencing the dispersion and concentration of pollutants in the atmosphere. By incorporating these factors into the modelling process, we can gain a more comprehensive understanding of the complex interactions between meteorology and air quality. These features can provide valuable insights into the underlying environmental conditions that contribute to air pollution levels. For example, high temperatures and low wind speeds can lead to the stagnation of pollutants, while rain can help to cleanse the atmosphere by removing particulate matter.

Additionally, leveraging advancements in remote sensing technology, such as satellite data or ground-based sensors, can provide real-time information on air quality and meteorological conditions. Integrating these data sources into the modelling framework can enable more dynamic and responsive air quality predictions, allowing for better management of air pollution and its impacts on public health and the environment.

REFERENCES

[1] Kothandaraman, N. Praveena, K. Varadarajkumar, B. Madhav Rao, Dharmesh Dhabliya, Shivaprasad Satla and Worku Abera (2022) Intelligent Forecasting of Air Quality and Pollution Prediction Using Machine Learning. https://www.hindawi.com/journals/ast/2022/5086622/

[2] Madhuri VM, Samyama Gunjal GH, Savitha Kamalapurkar (2020) Air Pollution Prediction Using Machine Learning Supervised Learning Approach https://www.ijstr.org/final-print/apr2020/Air-Pollution-Prediction-Using-Machine-Learning-Supervised-Learning-Approach.pdf .

[3] Yves Rybarczyk and Rasa Zalakeviciute (2021) Assessing the COVID-19 Impact on Air Quality: A Machine Learning Approach https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2020GL091202

[4] Doreswamy, Harishkumar K S, Yogesh KM, Ibrahim Gad (2020) Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models https://www.sciencedirect.com/science/article/pii/S1877050920312060?via%3Dihub

[5] Ling Qing (2023) PM2.5 Concentration Prediction Using GRA-GRU Network in Air Monitoring https://www.mdpi.com/2071-1050/15/3/1973

[6] Nilesh N. Maltare, Safvan Vahora (2023) Air Quality Index prediction using machine learning for Ahmedabad city https://www.sciencedirect.com/science/article/pii/S277250812300011X#:~:text=The%20Long%20Short%2DTerm%20Memory,Navares%20and%20Aznarte%2C%202020

[7] Ruiyun Yu, Yu Yang, Leyou Yang, Guangjie Han, and Oguti Ann Move (2016) RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems https://www.mdpi.com/1424-8220/16/1/86

[8] Nahar K, Ottom MA, Alshibli F, Shquier MA (2020) Air quality index using machine learning—a jordan case study. COMPUSOFT, Int J Adv Comput Technol 9(9):3831–3840 https://www.researchgate.net/publication/344438674_AIR_QUALITY_INDEX_USING_MACHINE_LEARNING_-A_JORDAN_CASE_STUDY

[9] Varsha Gopalakrishnan (2021) Hyperlocal Air Quality Prediction using Machine Learning https://towardsdatascience.com/hyperlocal-air-quality-prediction-using-machine-learning-ed3a661b9a71

[10] Mauro Castelli, Fabiana Martins Clemente, Ale˘s Popovi˘c, Sara Silva, and Leonardo Vanneschi (2020) A Machine Learning Approach to Predict Air Quality in California https://www.hindawi.com/journals/complexity/2020/8049504/

[11] Sanjeev D (2021) Implementation of machine learning algorithms for analysis and prediction of air quality. Int. J. Eng. Res. Technol. 10(3):533–538 https://www.ijert.org/implementation-of-machine-learning-algorithms-for-analysis-and-prediction-of-air-quality

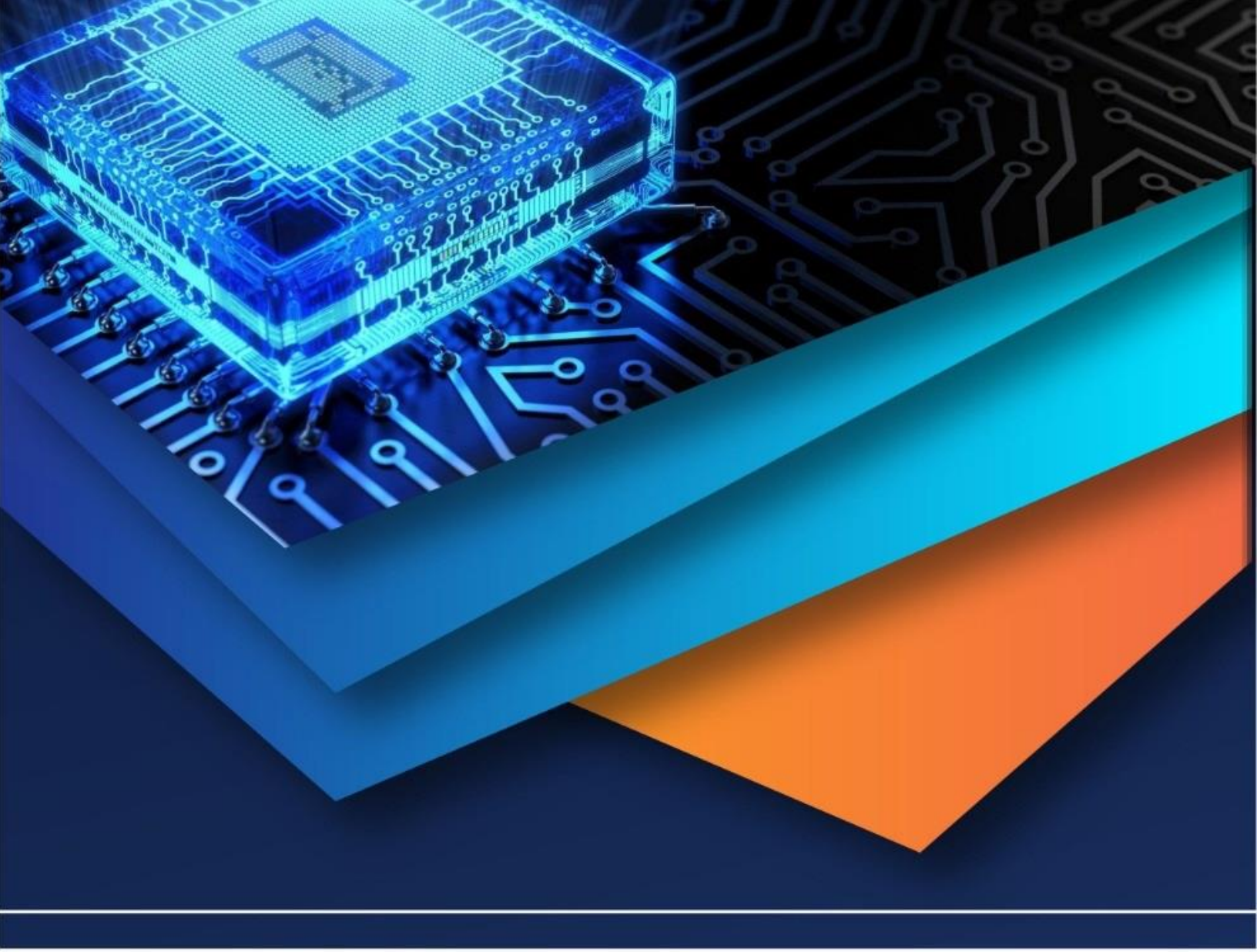
[12] Ghufran Isam Drewil , Riyadh Jabbar Al-Bahadili(2022) Air pollution prediction using LSTM deep learning and metaheuristics algorithms https://www.sciencedirect.com/science/article/pii/S2665917422001805

[13] JingyangWang, Xiaolei Li, Lukai Jin, Jiazheng Li, Qiuhong Sun1 & HaiyaoWang (2022) An air quality index prediction model based on CNN-ILSTM https://www.nature.com/articles/s41598-022-12355-6

[14] Tanisha Madan, Shrddha Sagar, DR. deepali Virmani(2020) Air Quality Prediction using Machine Learning Algorithms https://www.researchgate.net/publication/349802397_Air_Quality_Prediction_using_Machine_Learning_Algorithms_-A_Review

[15] Abdellatif Bekkar, Badr Hssina, Samira Douzi and Khadija Douzi (2021) Air-pollution prediction in smart city, deep learning approach https://link.springer.com/content/pdf/10.1186/s40537-021-00548-1.pdf

[16] Mr. Koteswara Rao Dasari, Dr. Srinivasa Rao Pendela, Mr.Nageswara Rao Itikala,, Dr. Sai Prasad Padavala, Dr. Kiran Sree Pokkuluri Hybrid Machine Learning Approach for Twitter Sentiment Analysis



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY